TRI AUTOMATIQUE SUR CRITÈRES MORPHOSYNTAXIQUES DE L'ANCIENNE LANGUE

Xavier-Laurent Salvador

TRI AUTOMATIQUE SUR CRITÈRES MORPHOSYNTAXIQUES DE L'ANCIENNE LANGUE

Xavier-Laurent SALVADOR (LDI)

Le présent article illustre les travaux du groupe d'études consacré à l'ancien français né dans le cadre de la réflexion de l'équipe LDI autour des problèmes de résolution automatique des problèmes linguistiques. Nous voudrions particulièrement présenter les prémices théoriques qui fondent notre travail de création de ressources au sein de la base G.R.A.A.L consacré à l'étude et au traitement automatique de l'ancien français.

Imaginons ainsi que nous proposions de mettre les capacités de calcul des ordinateurs contemporains au service non pas d'une application phénoménologique immédiate – étiqueter, trier, reconnaître, résoudre – mais plutôt d'un modèle philosophique qui chercherait à produire la langue dans son ensemble. Autrement dit plutôt que de concevoir les textes réalisés par les sujets parlants du XIIe siècle comme autant de corps étrangers auxquels se confrontent les analyseurs morphosyntaxiques pour en saisir les trois dimensions que sont la nature, la construction et les sens, il faudrait opérer une révolution du point de vue autour de l'objet. Ne pourrait-on pas imaginer que le corpus, loin d'être extérieur à la machine, pourrait être considéré comme l'un des chemins possibles auquel elle pourrait aboutir pour peu qu'on lui fournisse les moyens de produire de la langue ? Dès lors, toutes les applications mécaniques liées à l'identification, à l'analyse et à la reconnaissance du sens relèveraient du traitement de la coïncidence: la machine appréhenderait le texte en fonction de règles internes – sa propre perception des règles qui modélisent le langage – et non pas en fonction de critères qui lui sont artificiellement imposés par une dynamique externe généralement liée aux contraintes du corpus par l'ingénieur aux manettes. La machine verrait dans le texte un produit dont elle parierait à la base qu'il appartient à la sphère des textes qu'elle serait elle-même capable de produire. En un mot, que le texte soumis à son interprétation et elle-même appartiennent au même bain linguistique. Toutefois, nous entendons déjà les premières objections: à la manière des singes de la nouvelle de Borgès, l'ordinateur peut tout dire mais une chose est sûre: il ignore qu'il dit. Seul le regard de l'homme peut discriminer d'un coup d'œil le texte du charabia. Autrement dit, toute tentative de modélisation informatique du comportement linguistique se heurte à un invariant fondamental: le sentiment de la langue. Et paradoxalement, tout projet similaire se fixe plus ou moins consciemment de reproduire mécaniquement ce phénomène remarquable qui permet à un lecteur même le moins cultivé de combler instinctivement les lacunes morphologiques, syntaxiques et sémantiques pour saisir le sens général d'un texte, là où la machine la plus puissante s'épuise.

Il y a là un paradoxe maintes fois constaté que nous ne prétendons pas résoudre immédiatement, mais il y a là également source de réflexions sur une méthode d'approche du traitement automatique de la langue. Il s'agirait d'aborder le problème du traitement du corpus soumis à l'analyse non pas comme une fin en soi, mais comme une conséquence secondaire de la mise en place d'un protocole qui n'aurait pas pour but de traiter, mais de *produire* une langue. Toute une langue. Et lorsque nous disons « produire » une langue, nous entendons bien générer toutes les dimensions envisageables de la langue qui font la communication, du signifiant au discours. Une première étape importante pour

la mise en œuvre d'un tel procédé consiste avant tout à éprouver ce procédé dans un protocole de prédiction des formes de la langue afin de constituer une ressource exploitable *ad libidum* dans le cadre des applications décrites.
Le choix de l'ancienne langue, enfin – et nous conclurons ainsi notre propos liminaire – semble offrir de nombreux avantages pour l'évaluation de la pertinence des résultats observés. En effet, l'ancienne langue étant un état, elle offre deux conditions extrêmement intéressantes dans le cadre que nous nous sommes fixé: elle est toujours suffisamment éloignée des sujets parlants contemporains pour offrir les conditions du traitement diglosse mais suffisamment proche pour offrir à tout instant les moyens du recours à l'intuition de la langue sans recourir à l'intervention du bilingue. Autrement dit, et si l'on nous autorise la force de l'image pour illustrer notre propos, en se plaçant du point de vue interne nous pouvons considérer que l'ancienne langue – et particulièrement l'ancien français – constitue un espace clos mais non limité dont l'horizon est lointainement fixé par l'horizon toujours s'échappant de la modernité ; en nous plaçant du point de vue externe, l'ancienne langue est un corps fermé doté de ses règles et de ses principes dont la limite est posée par le seuil d'interprétation par les contemporains. Cet aspect fuyant et malgré tout observable du phénomène « ancienne langue » suffit à créer un objet de science passionnant et dont on imagine qu'il est susceptible de produire des résultats riches d'enseignements pour des domaines actuels. La résolution des problèmes de traduction, de reconnaissance et d'analyse des textes constitués par le corpus des textes en ancien français constitue donc une première étape dont nous voudrions exposer ici les prémices et les premiers résultats.

Modèle théorique de la démarche inductive

Pour construire la ressource G.R.A.AL, nous partons du principe qu'il faut imiter schématiquement les étapes cognitives qui président à l'élaboration par l'intelligence d'un vocabulaire. Autrement dit, il convient dans un premier temps de se poser la question suivante : si je devais épuiser intellectuellement toutes les formes qui constituent *ma* langue, comment mon intelligence s'y prendrait-elle pour coucher sur le papier tous les mots de la langue ? En imaginant qu'un esprit remarquablement intelligent puisse être mobilisé pour une telle épreuve, nous pouvons imaginer que sa démarche consisterait à essayer dans un premier temps de convoquer tous les mots de la langue qui font un lexique, puis qu'il les ferait varier en genre, en flexion et en nombre selon les règles propres par nature à chacun des mots ainsi convoqués. Devant l'ampleur de la tâche, on imagine aisément qu'au bonheur d'avoir rappelé le mot « ami » s'ajoute celui de pouvoir faire défiler les mots de la même famille: « amie », « amis », « amitiés » puis du même champ notionnel qui eux-mêmes appellent leurs propres réseaux: « copain », « copine », « copinage », « copiner ». Enfin, chacun de ces mots raccroche à la liste l'ensemble des éléments de leur paradigme: « copinent », « copinons ». Ce travail d'enrichissement du lexique serait enfin incomplet, et nous n'aurions pas une représentation exacte de la topographie linguistique de l'intelligence de notre cobaye si nous le laissions partir sans avoir épuisé la quantité infinie des variantes du même terme qu'il connaît. Il va de soi que « ami » et « copain » travaillent ensemble, de même que « pote » car ce dernier quoique d'un autre registre, appartient bien à la langue. Mais, pour peu que notre ami soit d'origine picarde, peut-être se souviendra-t-il que [kopẽ] et [kopɥ̃] sont un seul et même mot, même si sa grand-mère ne connaissait que le second. De même, peut-être se rappellera-t-il que son [kjẽ] accompagne encore son père à la pêche le dimanche. Autrement dit, il serait important dans le cas qui nous préoccupe de ne pas oublier que les variations sémantiques s'accompagnent en langue de variations phonétiques liées à la morphologie des mots eux-mêmes, mais également de variations phonologiques qui ne sont en rien anecdotiques. Ce n'est qu'au prix d'un tel effort que nous pouvons à un moment imaginer avoir parcouru et topographié *toute* la langue. Ce qui demeure notre objectif principal dans un premier temps.
Considérons maintenant l'épreuve à rebours: il s'agit de remplir mécaniquement la mémoire d'un

ordinateur de l'ensemble des données de la langue. Et, étant entendu que nous ne pouvons atteler un esprit d'homme à la tâche de consigner par écrit toutes les formes de la langue, il s'agit également de remplir cette mémoire de mécanique façon. Le choix de l'ancien français étant entendu pour l'objet, deux voies s'ouvrent devant nous:

- dans un premier temps, en vertu des règles que nous rappelions en introduction, nous pourrions considérer que l'ancien français étant un ensemble clos, une démarche productive consisterait à numériser l'ensemble de la production d'une époque afin d'en reconstituer le stock lexical. Les règles serviraient mécaniquement à rebours pour reconstruire les entrées dans la langue. Il s'agit d'une démarche déductive qui n'est pas dénuée de fondements, mais dont la capacité prédictive est purement contextuelle. Qui plus est, une telle approche interdit dans le long terme de pouvoir progresser, à savoir faire reproduire par l'ordinateur le comportement spécifique de l'intuition de la langue. Pour efficace qu'elle puisse sembler, une telle approche assigne à toute entreprise l'analyse de la langue comme unique objectif; or nous souhaiterions que l'analyse de la langue soit un résultat implicite du fonctionnement de notre théorie.
- Dans un second temps, une approche inductive plus conforme à nos souhaits consiste à partir des vocabulaires de la langue et d'y appliquer de manière automatique les règles de variations morphologiques, phonétiques et phonologiques décrites par les grammaires pour reconstituer artificiellement l'ensemble des mots de la langue.

Cette dernière démarche, qui rencontre particulièrement les attentes de notre conception de l'objet, n'est pas sans poser problème.

- En effet, dans le cas de l'ancien français, la première piste que nous pourrions être tentés de proposer consiste à appliquer à cette fin les règles que l'on utilise généralement pour parler de l'ancien français, à savoir la description philologique du français en diachronie. Il s'agirait donc de reproduire à partir d'un dictionnaire des étymologies l'ensemble des règles qui ont fait évoluer le français jusqu'à une époque donnée. Il s'agirait alors de saisir l'ensemble des étymons canoniquement retenus pour chacun des mots de la langue et d'y associer l'ensemble des règles ayant conditionné l'évolution phonétique des mots. Toutefois cette approche, outre qu'elle est extrêmement pesante, présente deux inconvénients majeurs : d'une part, elle ne permet pas de résoudre les filiations pour un ensemble de mots dont l'étymon demeure inconnu et d'autre part, nombre des règles ayant conditionné le comportement exceptionnel de certaines formes sont liées à la perception sémantique de celle-là, comportement que l'ordinateur aujourd'hui est incapable de reproduire. Un aspect analytique de l'évolution morphologique des paradigmes morphologiques intègre en diachronie une composante fondamentale : la dimension sémantique. Cette dernière conditionne en partie l'évolution de certaines formes en vertu de la perception en synchronie par le sujet parlant du sens étymologique d'une forme. Ainsi, par image, combien de gens aujourd'hui sentent-ils que le mot « *canicule* » est dérivé du mot « *chien* » ? Et bien, selon la perception qu'ont eue les hommes à différentes étapes de l'histoire de la langue, certains mots ont connu des évolutions atypiques par rapport au modèle attendu. Il existe par exemple une règle en philologie selon laquelle *dans les verbes composés* latins (ce sont les verbes qui nous intéressent au premier chef, mais nous pourrions multiplier les exemples), l'accentuation du latin classique s'est conservée « lorsque le sentiment de la composition était perçu[1] » mais a été modifiée dans la forme dérivée en français si le sentiment de la composition ne l'était plus. Cette alternative a produit des résultats phonétiques très divers : comment modéliser informatiquement le sentiment de la composition du mot ? Une évidence s'impose : il apparaît clairement que le rapport que nous entretenons avec la langue dans le cadre du projet S.A.C.R.E.G.R.A.A.L[2]. ne peut pas intégrer la dimension diachronique du français conçue comme la modélisation des structures ayant évolué. De

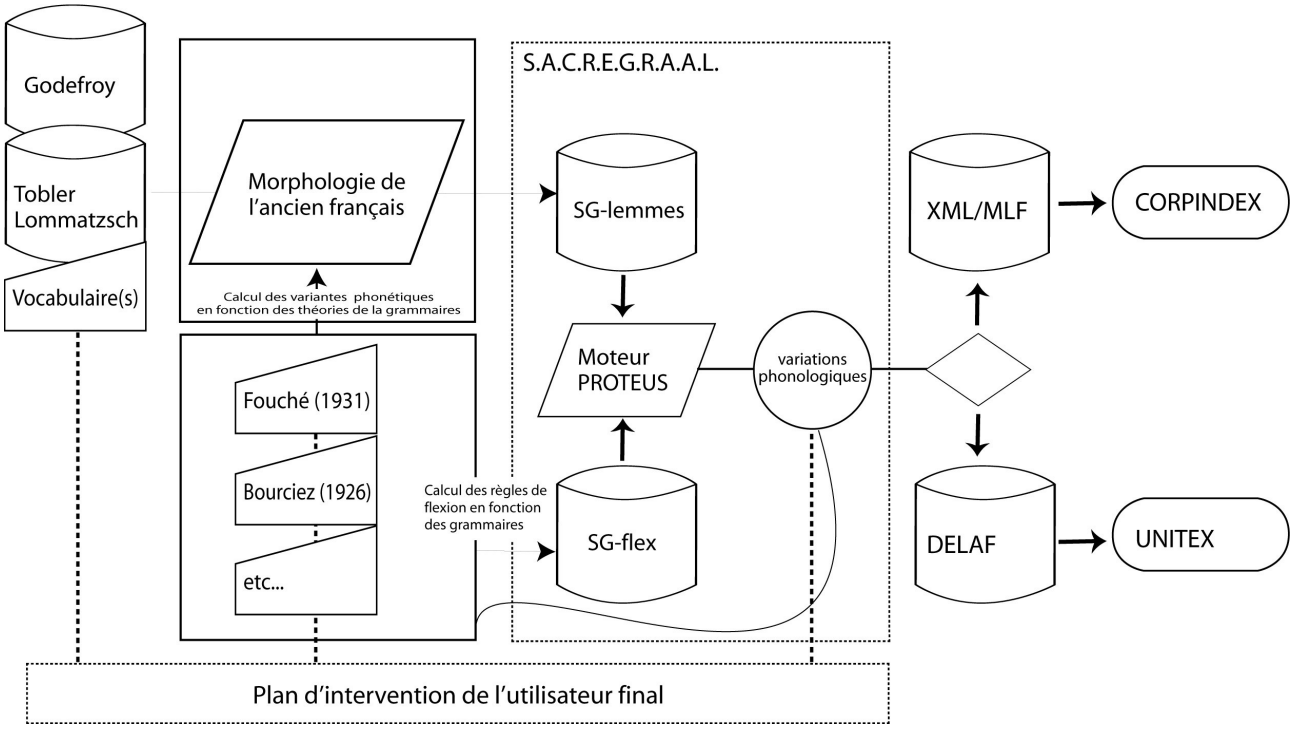
[1] Fouché P., *Morphologie historique du français...* ouv. cit. p. 4.

plus, il n'est pas non plus envisageable de prendre pour point de départ un hypothétique étymon/lemme censé rendre compte, d'un point de vue morphologique, de l'histoire d'un mot. Si cette solution peut être envisageable pour une faible proportion de mots, elle ne permet pas de gérer les changements sémantiques ayant altéré la forme ni ne permet d'intégrer le nombre notable de mots ayant un étymon non attesté ou reconstruit dans une langue non écrite ? Et que ferions-nous pour une forme comme « chétif » (f.m.) dont il faudrait définitivement affirmer que l'étymon est soit « *captivos* » (forme attestée en latin classique mais incohérente avec le résultat français; [kaptiuos] ne devrait pas avoir évolué sous la forme « chétif »); soit « **cactos* » (forme gallo-romaine jamais attestée en discours, mais cohérente avec la théorie philologique)[3].

– Une seconde approche que nous avons mise en place adopte un point de vue en synchronie sur la langue et lève ces résistances: en partant des vocabulaires fabuleux de l'ancienne langue produits par des érudits qui ont tout lu, nous appliquons les règles de variations morphologiques, phonétiques et phonologiques qui sont décrites par les grammaires philologiques comme le résultat de l'évolution de la langue. Nous nous réservons ainsi la possibilité de pouvoir filtrer le résultat de notre ressource en fonction de l'état de langue souhaité par un utilisateur final selon la période ou le lieu, en fonction des règles courantes du français standard ou de règles spécifiques liées à un usage localisé. Nous nous réservons même ainsi la possibilité de pouvoir étendre la ressource aux bases du français moderne, de sorte que cette dernière démarche nous offrira la possibilité de pouvoir prolonger notre enquête à d'autres secteurs du fonctionnement global de la langue mécanique.

Illustration de l'interaction avec l'utilisateur final

Le schéma suivant voudrait donc donner corps aux différentes étapes que nous venons de décrire:



[2] Système Avancé de Conception de Ressources Électroniques pour la Gestion et la Reconnaissance Automatique de l'Ancienne Langue.

[3] Pour d'autres exemples d'incompatibilités, nous renvoyons entre autre à Chambon J.-P. , « Cas d'étymologie double dans le FEW », *Travaux de linguistique et de littérature*, 27, 1989 pp. 151-179.

Dans un premier temps, le moteur GRAAL reçoit en entrée les bases des deux dictionnaires de l'ancien français standard: le godefroy et le Tobler-Lomatzsch. Le premier dictionnaire est l'outil de référence des philologues français. Le lexique constitué quant à lui par Tobler, Lommatszch et Christman dans la version révisée par les éditeurs de la version électronique intègre dans les entrées plus de 30 000 variantes renvoyant à d'autres entrées. Autant dire que la conjugaison de ces deux outils forme un aperçu large du potentiel de la langue médiévale. Nous avons fait le choix de laisser les entrées des deux dictionnaires, même lorsque celles-là semblent redondantes. Pourquoi ? Dans la perspective de devoir lever les ambiguïtés après projection de la ressource, il ne nous semble pas anodin de pouvoir déterminer si une forme est présente dans les deux lexiques, ou bien dans un seul des deux et de savoir dans lequel. Nous avons donc ajouté, dans la nomenclature de la base de données, une étiquette « Dictionnaire d'origine ». SG-lemme enfin, ce sont en quelques chiffres et à l'heure où nous nous employons à en illustrer les perspectives d'application, source TL : 76908 entrées; source GD : 55932; nombre de lemmes : 132866 et nombre d'hyperlemmes : 82811.

Dans un second temps, nous représentons le filtre morphologique qui s'appuie sur la symbolisation des règles de variations énoncées par les grammaires de référence. C'est à ce stade que nous produisons automatiquement les déclinaisons complètes des déterminants, des pronoms, des adjectifs. C'est à ce stade que se règlent les formations du pluriel, mais aussi le traitement de certaines particularités morphologiques comme l'apocope. Ainsi, de l'entrée « cheval » présente dans les deux dictionnaires, nous produisons « chevals », forme du pluriel canonique. Mais nous ne produisons pas « chevaus » ni « chevaux », les formes attestées qui sont des accidents morphologiques par rapport à la forme orthologique[4]. L'ensemble de ce répertoire constitue SG-lemmes, la base fondamentale des mots de l'ancienne langue.

Ce répertoire est destiné au traitement des flexions dans le corps de Proteus, un automate mis au point au sein du LDI, qui a pour vocation de décrire et appliquer des règles de flexion au moyen d'un langage propre. Nous renvoyons à la communication de notre collègue pour plus de détails sur le fonctionnement spécifique de Proteus, mais quoi qu'il en soit, à ce stade, l'automate dispose de deux réservoirs dédiés à la formation de la langue: SG-lemmes, base lexicale et SG-flex, un ensemble de règles de flexions destinées à traiter la base initiale.

Enfin, au sortir de l'automate nous adaptons un moteur de traitement des variations phonologiques qui est une adaptation des règles décrites par les grammaires. L'ensemble produit de la norme, c'est-à-dire un outil d'étiquetage xml adaptable à toute situation d'analyse de corpus.

Le principal obstacle que nous rencontrons à ce stade de l'élaboration est lié à la double contrainte spécifique de l'ancienne langue, considérée comme un ensemble homogène par nos contemporains mais dont la réalité est incontestable au regard du nombre de siècles et de dialectes que le terme « ancien français » recouvre. Il est donc nécessaire d'ajouter deux points importants:

- dans un premier temps, le traitement des données décrites par la grammaire n'ignore pas les spécificités de l'ancienne langue ni l'ampleur du spectre diachronique de l'objet. Pour résoudre le problème lié à la double contrainte de l'unité de l'objet et de la grande variété spatiales et temporelles qu'il recouvre, nous avons mis au point un système de marqueurs au sein même de la base de données qui permet d'affiner la description de l'hyperlemmes en fonction du lieu et de la date.
- Dans un second temps, nous sommes parfaitement conscients que notre connaissance est à ce jour partielle et ne saurait prétendre tout dire de tous les dialectes à toutes les dates. Toutefois, nous prétendons d'une part avoir décrit un état de langue standard susceptible d'être mis en exercice à des fins pédagogiques. D'autre part, nous souhaiterions que GRAAL soit un outil exploitable par les philologues à leurs fins propres.

Le plan de l'utilisateur final caractérise les moyens d'intervention et de guidage de tout utilisateur aux différents points d'entrée de la chaîne de production. Nous repérons trois points d'intervention

[4] Sur cette notion, voir notamment J.-C. Chevalier et M.-F. Delport, *L'Horlogerie de saint Jérôme: problèmes linguistiques de la traduction*, Paris, L'Harmattan, 1995, pp. 30 *et sqq.*

qui correspondent dans la réalité matérielle de l'objet à trois modules externes susceptibles d'interagir avec la console. Il s'agit du vocabulaire d'entrée, des règles de variations morphologiques et des règles de variation phonétiques. Tout utilisateur, à court terme, pourra intervenir sur chaque élément indépendamment des deux autres. Ainsi, un chercheur qui voudrait produire une norme d'étiquetage automatique pour des besoins spécifiques liés à la description qu'il est en train de construire des règles de l'ancien gascon au XIIe siècle pourra saisir – dans le respect des normes que nous sommes en train de publier – le vocabulaire spécifique rencontré à la lecture de ses archives; intervenir sur les règles de conjugaison propres à ce dialecte en saisissant les désinences et les modèles de variation; faire varier les formes en fonction des caractéristiques de ce dialecte sur la base d'un dialogue avec la machine fondé sur le parcours des différentes possibilités: voyelles/consonnes/diphtongues.

Nous voyons donc qu'à moyen terme l'outil GRAAL est destiné à devenir un assistant à la mise en forme et à la description de la langue. L'analyse et l'exercice sur corpus seront des moyens d'enrichir la base de connaissances – car c'est bien d'une base de connaissances dont il s'agit – qui permettra de valider les hypothèses émises par la machine.

On nous objectera peut être que la surgénération inévitablement entraînée par le choix du modèle inductif entraîne une série enchaînée de lourdeurs qui ralentissent le traitement et qui, à terme, pourraient en empêcher la bonne marche. C'est effectivement là le problème majeur que nous rencontrons dans la mise en ligne de l'application aujourd'hui: les calculateurs qui appliquent la « force brute » aux modèles linguistiques prennent encore du temps à déployer toute leur activité « à la volée ». Là encore, deux pistes s'offrent à nous pour pallier ce manque d'efficacité qui à terme pourrait devenir un obstacle incontournable: dans un premier temps, parier sur l'adaptation des calculateurs. Au rythme de développement des machines aujourd'hui, on imagine sans peine que la puissance ne sera bientôt plus un argument. La seconde piste ouvre également une série de chantiers qui devraient se révéler passionnants: l'apprentissage. En effet, rien n'interdit de penser aujourd'hui que les différents exercices de la machine sur des corpora traités et vérifiés de main d'homme pourront à terme devenir autant de moyens pour la machine d'apprendre, par calculs statistiques par exemple, la nature vraie ou fausse des résultats affichés dans le cadre de procédures d'étiquetage automatiques. Alors les vocabulaires se spécialiseront et rien n'interdit enfin d'imaginer, mais il faudra beaucoup travailler, appliquer les mêmes méthodes aux règles de syntaxe qui président au destin de la proposition et du groupe nominal, puis au discours. Et d'étendre enfin ce travail à toutes les anciennes langues romanes.

Avancées et perspectives

Pour l'heure, examinons tout d'abord ce que fait GRAAL. Ainsi, mettons-nous dans l'optique d'un professeur chargé de la question de grammaire pour la préparation du concours de l'agrégation. Ce dernier se propose de travailler sur les temps de l'indicatif dans le corpus au programme - *Le Jeu de Robin et Marion*[5] - et pose la question suivante: « Quels sont les verbes conjugués aux différents temps de l'indicatif ? ». Notre ressource, après traitement automatique du texte au programme de l'agrégation, est d'ores et déjà capable de fournir un relevé convaincant.

Dans un premier temps, nous récupérons toutes les formes de verbe du premier groupe en fonction de la désinence puis, après élimination de l'apparat d'étiquetage, chacune est testée selon la nature de sa forme. Nous isolons six variantes de verbes de base I qui présente une particularité morphologique intéressante puisqu'il s'agit des verbes qui prennent un « e » de désinence en P1 de manière régulière :

-b(i)er
-p(i)er
-g/v/b/c/k/q/qu/(i)er
-g/v/b/c/t/d/k/q + r(i)er

[5] *Idem*.

-c/g/b/k/q+l(i)er
-gn(i)er

Dans un second temps, pour faciliter le calcul des paradigmes de la conjugaison, nous isolons au moyen d'un script simple la voyelle tonique de chacune des formes de sorte que nous isolons différentes natures de verbes du premier groupe *distingués en vertu de leur propriétés*. Ainsi, nous obtenons

base en « e/ë » avec ou sans entrave (« e, o, eu, ie, oe, ei, oi, ii, ee »)
base en « o/ö » avec ou sans entrave (« o, u, oi, oe, eo, ue, ou, eu, ei, ueu, uou »)
base en « i/ï » avec ou sans entrave (« i, ei, oi, e, i »)
base en « uu/ü » avec ou sans entrave (« o, u, oi, oe, eo, ue, ou, eu, ei, ueu, uou »)
base en « a » avec ou sans entrave (« a, ai, ei, e »)
base en « ou » (« ou,o,u,eu, ue, ueu, uou »)
base en « au » et base en « ao » (« au, o, u, uau, ueu, eu, aiu, uao, ueo, aio, ado, adu »)
base en « eu » (« eu, o, u,eu, e, ueu, uou »)
base en « ue » (« ue, e , o, ieu, uieu, i »)
base en « oi » (« e, oe, oi, ei, i »)
base en « ei » (« ei, oi, oe e, i »)
base en « ii » (« ii, ei, oi, i, ie, illi, ill, lli, li, il »)
base en « ai » (« ai, ei, a, e, ailli, alli, ali, ail, eil, eilli, eill, elli, eli »)
base en « oo » (« oo, ou, eo, ao, au, eu, ue, ueu, uou, uau, odo, od »)
base en « ua » (« a, ai, ei, e, ua, uai, uei, ue »)
base en « iu » (« iu, iou, ieu, illiu, illiou, illieu, illou, illeu, illu, liu, lio, lieu »)
base en « ui » (« ui, uoi, uei, ueu, eu, ei, oi, uilli, ulli, uli, uill »)
bases avec entraves autres que mentionnées précédemment (Il faut rajouter une forme identique, mais on enlève la consonne de l'entrave)

Dans un troisième temps, nous envisageons une série de formes qui prennent en compte les variations phonétiques dialectales, hasardeuses, historiques de la langue. Nous calculons ainsi, en fonction des règles de la grammaire, différentes variantes selon qu'il y a une liquide immédiatement après la tonique, une nasale, une double nasale, une occlusive sourde, une chuintante, une sifflante. Enfin, nous appliquons de manière indifférenciée à toute la base de données un calcul de variantes affectant tout le mot, et plus simplement l'entourage direct de la voyelle tonique.
Et cela commence à produire certains résultats:
Le moteur distingue les verbes de P6:

{verbes_finissant_par_ons} : Huars musera , et chil doi autre corneront. Or (*ostons*) ains ches coses dont .
{verbes_finissant_par_ons} : Mais nous (*arons*) anchois balé , entre nous deus , car bien balons.
{verbes_finissant_par_ons} : arons anchois balé , entre nous deus , car bien (*balons*) [...]
Les verbes de P4 et P5, ou verbes pluriel:
{verbes_présent_pluriel} : Robins : or (*esgardons*) leur destinee , par amours , si nous embuissons [...]
{verbes_présent_pluriel} : Marote : et encore ! (*esgardez*) comme est reveleus ! robin , tu ies moult corageus [...]
Les différents temps:
{verbes_futur_singulier} : Mais ore faisons feste de nous. Robins : (*serai*) je drois ou a genous ?
{verbes_futur_singulier} : Robins : et jou l' otroi ; je (*serai*) chi , lès ton costé.
ou les verbes à l'imparfait:
{verbes_imparfait_1pers} : a quoi pensés vous ? Gautiers : certes , je (*pensoie*) a Robin.

Les requêtes croisées permettent ainsi d'obtenir très rapidement des corpus étiquetés et manipulables à volonté. Il reste de nombreux obstacles et de nombreuses sources de confusions à résoudre. Nous pensons notamment aux ambiguïtés de reconnaissance prétérite/futur où la proximité des désinences pour les verbes du premier groupe est grande. En effet, le moteur aujourd'hui propose une solution fausse:

{verbes_futur_singulier} : sain et une grant pieche de pain qu' il m' (aporta) ore a prangiere .

Il s'agit là d'un problème classique en traitement automatique des langues lié à l'ambiguïté inhérente de la langue, amplifié ici par la présence de nombreuses variantes. Il nous faudra du temps pour trouver les équilibrages satisfaisants en éliminant les aberrations pour pouvoir utiliser des

techniques classiques de levée d'ambiguïté, notre idée étant de focaliser sur une ressource linguistique – le dictionnaire – la plus complète et précise possible. Toutefois, la voie que nous avons choisie semble prometteuse. Après quelques mois d'élaboration du projet, la production automatique de documents pédagogiques semble à portée de main. Par ailleurs, nous sommes d'ores et déjà en mesure de produire un outil d'entraînement efficace pour l'analyse de l'ancienne langue. Au-delà, nous pouvons déjà offrir aux chercheurs la possibilité d'inclure les règles de génération et de variation en fonction des besoins de leur recherche. L'application pédagogique d'un tel outil est infinie, depuis la création automatique d'exercices de repérages pour de jeunes philologues, à la mise en place de requêtes complexes sur le mode des questions des concours.

Conclusion

Le tri automatique sur critères morphosyntaxiques de l'ancienne langue est une première étape dans l'élaboration d'un outil susceptible d'intégrer la dimension historique de la langue du point de vue informatique. Le va-et-vient fondamental entre la démarche théorique, *forcément inductive*, et le travail pratique sur corpus permet la confrontation de l'automate à la réalité de langue et ouvre de nouvelles perspectives d'application au champ d'autres langues romanes.

Bibliographie

Andrieux N. et Baumgarner E., *Système morphologique de l'ancien français. A. Le verbe*, Paris, Klincksieck, 1983.

Bourciez E., *Précis historique de phonétique française*, 6e éd., Paris, 1926.

Brunot F., *Grammaire historique de la langue française*, Paris, 1886.

Chabaneau C., *Histoire et théorie de la conjugaison française,* Paris, 1878.

Chambon J.-P., « Les emprunts du français moderne aux dialectes et aux patois: une illusion d'optique en lexicologie française historique », *Lalies. Actes des sessions de linguistique et de littérature*, 17, 1997, pp. 33-53.

Fouché P., *Morphologie historique du français – Le Verbe*, Paris, Klincksieck, 1976.

Fouché P., *Le Verbe français*, Paris, 1931.

Doudet E., Méot-Bourquin V., *Adam le bossu, Le Jeu de la feuillée, Le Jeu de Robin et Marion, Jean Bodel, Le Jeu de saint Nicolas*, Neuilly-Sur-Seine, Atlande, coll. « Clefs Concours Lettres médiévales », 2008.

Heiden S. & A. Lavrentiev, « Ressources électroniques pour l'étude des textes médiévaux : approches et outils. » in *Revue française de linguistique appliquée*, IX(1), pp. 99-118, 2004.

Meyer-Lübke, *Historische Grammatik der franzoesischen Sprache,* Heidelberg, 1908.

Prévost S., Heiden S., « Etiquetage d'un corpus hétérogène de français médiéval : enjeux et modalités », 6 - 8 Octobre 2000, 1st Freiburg Workshop on Romance Corpus Linguistics, Freiburg, in : *Romanistische Korpuslingustik: Korpora und gesprochene Sprache, Romance Corpus Linguistics: Corpora and Spoken Language* p. 127-136, Pusch C. D., Raible W. (eds), Gunter Narr Verlag, Tübingen, 2002.

Stein A., « Étiquetage morphologique et lemmatisation de textes d'ancien français » - Kunstmann, Pierre et. al. (ed.): *Ancien et moyen français sur le Web: Enjeux méthodologiques et analyse du discours*, Ottawa: Les Éditions David, 2003, 273-284.

Skarup P., *Morphologie Synchronique de lancien français, Etudes romanes*, Museum Tusulanum, Copenhague, 1994.

Straka G. (ed.), *Les Dialectes de France du Moyen Âge et aujourd'hui:domaine d'oil et domaine franco-provençal*, Paris, Klincksieck, 1972, 478 p.

Walker Douglas C., *Old French Morphophonology. Studia Phonetica* 19, Didier, Ottawa, 1981.